三种数据挖掘算法在电子病历知识发现中的比较*

牟冬梅[1]　任　珂[2]
[1](吉林大学公共卫生学院　长春 130021)
[2](武汉大学信息管理学院　武汉 430072)

摘要:【目的】从异构的电子病历数据中发现疾病危险因素，为数据挖掘与知识发现提供借鉴。【方法】选取集各种结构为一身的临床电子病历数据，利用决策树、逻辑回归和神经网络三种数据挖掘算法分别建立疾病危险因素预测模型，对三种预测模型进行比较分析和统计学评价。【结果】决策树预测模型在查准率、召回率上高于逻辑回归和神经网络，在总体性能上决策树最优，但三者差别不大。【局限】未对电子病历属性进行优化选择。【结论】决策树在危险因素的发现与疾病的预测方面优于逻辑回归和神经网络。研究中建立基于数据挖掘算法的异构数据源知识发现框架，为今后领域知识发现和知识库构建以及数据挖掘算法的选择提供一定借鉴和参考。

关键词: 知识发现　电子病历　数据挖掘算法　预测模型
分类号: G202

1 引　言

随着大数据(Big Data)概念的提出及大数据时代的到来，情报学研究范畴已经明显呈现出大数据的典型特征[1]。大数据具有的数据量大、处理速度快、数据类型繁多和价值密度低这“4V”特征，为情报学提出新挑战，尤其大数据种类繁多、结构多样、质量参差不齐，情报学领域信息加工需要向数据清洗、规范集成和整合不断拓展。美国管理学家罗素·艾可构建了DIKW(Data-Information-Knowledge-Wisdom)体系[2-3]，Zeleny区分了DIKW体系中的各个元素[4]，CIO时代网对其内容与价值进行分析[5]，王曰芬认为文献计量法和内容分析法是实现DIKW转换的关键算法[6]。DIKW体系为情报学提供了巨大的发展空间，同时也指明情报学研究的目的和内涵，情报学需要在数据清洗的基础上，通过自然语言处理、概念映射等情报学方法进行数据标准化、规范化，再利用内容分析、科学计量分析、社会网络分析等多样化数据分析算法，通过数据挖掘提取内在的隐性知识，实现知识发现，为用户提供嵌入式的个性化精准化服务。

目前医疗数据是最为复杂的数据，最能体现大数据种类多、来源多、用途多的特征，本研究选取临床电子病历(Electronic Medical Record, EMR)数据，在情报学知识发现框架指导下，利用决策树、逻辑回归和神经网络等数据挖掘算法分别建立疾病的危险因素预测模型，并对三种预测模型进行评价。本研究规范情报学方法在医学领域知识发现的流程，探索从复杂的数据中找到知识之间有效关联及知识发现的最佳算法，为今后数据处理和知识发现提供一定借鉴和参考；另一方面，可以为临床医生的诊断提供数据支持，为疾病防控人员提供可视化依据，对妊高症“预防–诊断–治疗–预后”全过程提供科研数据支持；数据挖掘方法

通讯作者：任珂，ORCID：0000-0003-3366-1924，E-mail：lansexinghuo@163.com。

*本文系国家自然科学基金项目“嵌入式知识服务驱动下的领域多维知识库构建”(项目编号:71573102)和吉林大学大学生创新创业训练计划“基于数据挖掘算法的体检数据中脂肪肝危险因素相关性研究”(项目编号: 2015721054)的研究成果之一。

应用于疾病的危险因素研究，可以加强对医疗大数据信息的开发与利用。

2 基于数据挖掘算法的异构数据源知识发现框架

基于数据挖掘算法的异构数据源知识发现遵循科学领域逻辑框架内的知识发现研究[7]，在知识处理流程中关注数据规范，对不同来源的数据在异质领域本体融合基础上实现数据语义规范化，进而深入探讨主题模型、关联数据分析及机器学习等方法，是实现高效领域知识发现的一条必经之路，其流程主要有 4 步，如图 1 所示：

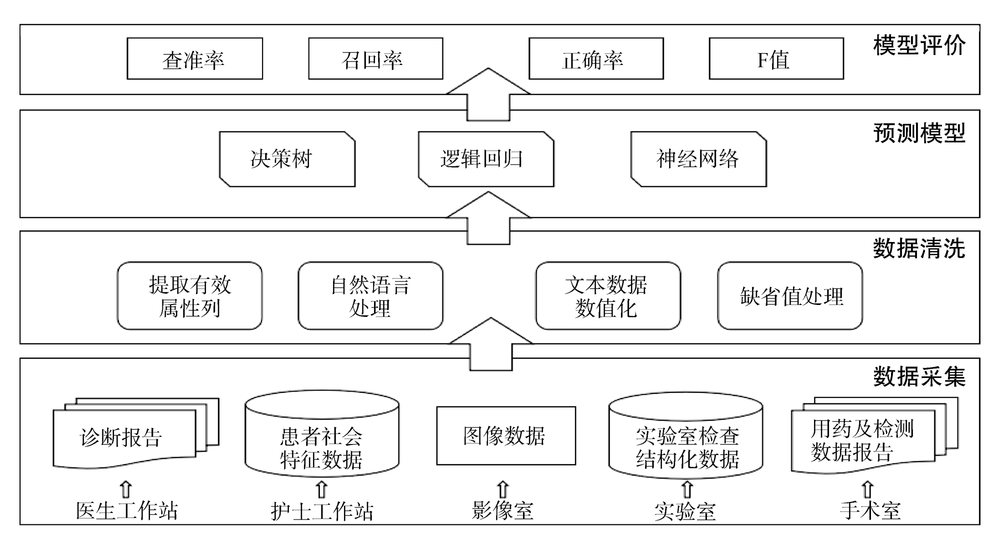


图 1　基于数据挖掘算法的异构数据源知识发现框架

(1) 利用数据库技术完成数据采集。涉及多个数据来源，如医生工作站的诊断报告、护士工作站的患者社会特征数据、影像室所保存的图像数据、实验室所保存的实验室检查结构化数据和手术室的用药及检测数据报告等，不同来源的数据呈现多种结构，将不同结构的数据进行结构化，存放在数据库中。

(2) 数据清洗。完成数据去标识化，数据类型的规范化，缺省值处理，自然语言处理[8]和语义标注。其关键技术为自然语言处理和语义标注。

(3) 预测模型构建。运用机器学习[9]中的监督学习方法[10]，进行疾病危险因素的预测模型构建。从大量多维度的数据中挖掘出有价值的情报，分析数据背后的知识。数据挖掘技术包括许多算法，按照训练的数据有无标签，分为监督学习算法、无监督学习算法和特殊算法，本研究应用开源软件R[10]建立数据挖掘模型。在R中通过运用决策树、逻辑回归和神经网络三种数据挖掘算法对相关数据分别进行处理分析，包括去掉缺省值、发现异常点、对数据进行唯一化处理和对相关的类目进行关联分析等，最终建立合理有效的数据挖掘模型，利用R软件的相关函数进行模型可视化展示，并通过模型对数据进行预测，从而得到有效的处理结果。

(4) 模型评价。对预测模型利用统计学方法进行评价，评价指标包括查准率、召回率、正确率和 F 值。

3 疾病危险因素预测模型构建

3.1 数据来源

研究数据来自长春市某三级甲等医院的电子病历，包含 2014 年 1 月 1 日至 2015 年 4 月 30 日就诊于该所医院的 31 443 名孕妇的就诊信息，由信息中心人员进行数据抽取建立 Excel 数据库(见图 2)。数据包括：病人的基本信息(科室、年龄、登记号、性别、民族、职业、文化程度、婚姻状况、收入)；生活和工作习惯信息(吸烟情况、饮酒情况、工作压力及精神压力)；病史信息(既往史、家族史)；常规体检数据(身高、体重)和实验室检查数据(收缩压、舒张压、总胆固醇、甘油三酯、低密度胆固醇、高密度胆固醇、空腹血糖、血红蛋白)；诊断结果。每名患者都严格按照医学诊断标准进行诊断，并且按照电子病历的格式，在既往史、家族史和诊断结果中用自然语言形式进行详细描述。

年龄	登记号	性别	民族	职业	文化程度	婚姻	吸烟	饮酒	既往史	身高cm	体重Kg
35	295405	女	汉族	-	大学或以上	已婚	无	无	否认肝炎、结核	160	80
29	245642	女	汉族	无	大学或以上	已婚	无	无	否认其他重大疾	156	78
29	245642	女	汉族	无		已婚	无	无	患者否认重大疾	156	77
32	157760	女	汉族	无	高中/中专	已婚	无	无	既往健康，无高	167	82
31	186432	女	汉族	职员	高中/中专	已婚	无	无	患者否认重大疾	153	59.5
34	122794	女	汉族	无	高中/中专	已婚	无	无	既往健康，无高	151	86
28	94276	女	汉族	技术干部		已婚	无	无	患者否认其他重	160	68
29	36740	女	汉族		大学或以上	已婚	无	无	健康，否认结核	155	74
34	44392	女	汉族	无业		已婚	无	无	患者因“异位妊	160	68
33	114062	女	汉族	无	大学或以上	已婚	无	无	否认重大疾病史	155	70
33	146330	女	汉族	职员	大学或以上	已婚	无	无	否认肝炎、结核	168	73
43	293197	女	汉族	无		离婚	无	无	良好	160	52
32	1019	女	汉族		高中/中专	已婚	无	无	否认肝炎结核病	160	81
30	42243	女	汉族	无		已婚	无	无	既往健康，无高	158	60
32	106048	女	汉族	职员	大学或以上	已婚	无	无	健康。否认肝炎	168	79
28	225887	女	汉族	职员	高中/中专	已婚	无	无	患者否认其他重	158	72
28	120532	女	汉族	无	初中	已婚	无	无	患者否认其他重	160	70
28	4929	女	汉族	无业	高中/中专	已婚	无	无	患者否认其他重	158	62

图 2　原始研究数据(部分)

3.2　数据清洗

(1) 提取有效属性列

由于数据中有些属性对于预测模型无影响或影响极小，加入分析可能会形成噪音(如入院日期、登记号等)，在提取有效属性列阶段，将噪声属性列去掉，保留有意义的属性列。研究中主要采用人工抽取的方式进行属性列提取，加大提取的准确性和有效性。

(2) 自然语言处理

对电子病历中既往史、家族史和入院诊断、出院诊断等自然语言描述的非结构化信息进行处理。首先进行二分类判别，对既往史、家族史以标点符号为分隔符进行数据提取，对分离出的疾病名称数据以及出入院诊断中的疾病名称进行概念映射，映射到统一医学语言系统(Unified Medical Language System, UMLS)下的国际疾病分类法 ICD10 中，方便之后数据挖掘模型对数据的有效识别。

(3) 文本数据数值化

在数据挖掘模型中，神经网络只能处理数值型变量，因此为了便于数据挖掘模型的建立，在本阶段将定性数据改为数值型变量。例如，在“婚姻状况”属性列中，设“离婚”为 1，“已婚”为 2，“未婚”为 3，“丧偶”为 4，“其他”为 5 等。

(4) 缺省值处理

由于电子病历记录不规范，存在病人记录不完整的现象，这些病人记录会影响最终模型的建立和挖掘，但由于这些缺省值并不多，因此使用 R 软件将含有这些缺省值的数据去掉，以呈现更好的挖掘效果。

通过上述步骤完成基本数据准备，使数据呈现可处理状态，也使数据库中的数据可以更加清晰简明地表述出来，最终得到 29 901 条数据，如图 3 所示:

年龄	婚姻	吸烟	饮酒	既往史	家族史	身高cm	体重Kg	收缩压mmHg	舒张压mmHg	总胆固醇	甘油三酯	低密度胆固醇	高密度胆固醇	空腹血糖	血红蛋白	编号	合并结果
28	2	2	2	2	2	160	84	110	70	3.41	4.97	0.88	0.97	10.99	83.00	30648	是
45	2	1	2	2	1	160	65	110	70	6.15	1.85	3.98	1.80	5.06	119.00	14172	否
31	2	2	2	1	2	171	101	110	70	6.47	4.22	4.07	1.56	5.63	67.00	11280	否
45	2	2	2	2	2	158	56	120	80	3.43	1.14	1.79	1.07	4.46	79.00	30615	否
28	2	2	2	2	2	168	72	90	60	5.68	2.65	3.55	1.60	4.87	133.00	18358	否
34	2	2	2	1	2	160	61	90	60	8.35	2.02	5.94	2.01	4.00	109.00	1664	否
29	2	2	2	2	2	168	65.5	110	80	6.00	1.08	2.34	2.61	7.54	94.00	15500	否
29	2	2	2	2	2	155	70	100	60	6.88	3.25	3.93	2.30	3.95	113.00	8015	否
28	2	2	2	2	2	162	87	130	80	4.57	4.45	1.81	1.87	6.02	76.00	26497	否
39	2	2	2	1	1	155	55	120	70	5.66	0.69	3.37	1.16	5.67	79.00	18622	否
27	2	2	2	2	2	152	61	110	70	6.77	2.17	3.55	2.20	4.57	85.00	8862	否
25	2	2	2	2	2	165	74	120	80	4.79	1.40	2.52	1.07	4.90	90.00	25259	否
21	2	1	2	2	2	155	55	110	70	6.19	2.59	2.76	2.32	4.93	105.00	13787	否
30	2	2	2	2	2	160	80	100	70	6.86	5.30	2.77	3.03	6.55	120.00	8170	否
43	2	2	2	2	2	158	56	120	80	4.92	4.06	2.17	1.04	5.31	87.00	24470	否
27	2	2	2	2	2	160	76	100	70	8.09	3.95	4.03	1.97	4.34	133.00	2277	否
31	2	2	2	1	2	155	61	110	70	4.98	3.25	2.48	1.85	4.70	97.00	24057	否

图 3　数据处理后的研究数据(部分)

3.3　妊娠高血压综合征的危险因素预测模型构建

针对挖掘妊娠高血压综合征(Pregnancy-induced Hypertension，简称妊高症)危险因素来进行上述算法的实证研究。在完成上述数据准备处理阶段的工作后，为了研究的一致性和严谨性，三种数据挖掘算法都应使用相同的训练集和测试集，将数据按照 7∶3 的比例分为训练集数据和测试集数据，选取 70%(即 22 010 条数据)的数据作为训练集，建立数据挖掘模型以及挖掘妊高症危险因素；剩余 30%(即 9 433 条数据)的数据作为测试集，用来测试算法性能。随后在 R 中对训练集和测试集数据的缺省值进行删除，最终结果为：训练集数据 20 940 条，测试集数据 8 961 条。

(1) 决策树模型

决策树作为一种监督学习方法，可以用于分类和

预测，在其树型结构中，每个节点和分支都具有一定的含义：决策树通过不断细化的分支(即分类标准)，将错综复杂的数据分为若干类型，用叶子节点对其进行表示，因此决策树可以对数据进行直观明确的分类。本研究采用 ID3 算法构建决策树模型。要构造尽可能小的决策树，关键在于选择合适的产生分支的属性。而 ID3 算法的核心正是通过采用信息增益的方式来选择能够最好地将样本分类的属性[12]。

设 $E = D_1 \times D_2 \times \cdots \times D_n$ 是 n 维有穷向量空间，其中 D_j 是有穷离散符号集，E 中的元素 $e = < v_1, v_2, \cdots, v_n >$ 为例子，其中 $v_j \in D_j$，j=1, 2, 3, …, n。设 s_1，s_2，…，s_m 是 E 的 m 个例子集。假设向量空间 E 中的这 m 个例子集的大小为 S_i，ID3 基于以下两个假设[13]：

(1) 在向量空间 E 上的一棵正确决策树对任意例子的分类概率同 E 中这 m 个例子的概率一致。

(2) 一棵决策树能对一个例子做出类别判断所需的熵为：

$$\mathrm{Entropy}(s_1, s_2, \ldots, s_m) = -\sum_{i=1}^{m} p_i \log_2(p_i) \tag{1}$$

其中，p_i 用 s_i / s 来估算。

如果以属性 A 作为决策树的根，A 具有 v 个值，它将 E 分成 v 个子集 $\{E_1, E_2, \cdots, E_v\}$，假设 E_i 中含有 S_i (i=1, 2, …, m)，那么子集 E_i 所需的期望信息是 E(A)。

$$\mathrm{Entropy}(A) = -\sum_{j=1}^{v} (s_{1j} + s_{2j} + \cdots + s_{mj}) / s \times \mathrm{Entropy}(s_{1j}, s_{2j}, \cdots, s_{mj}) \tag{2}$$

因此，以属性 A 为根的信息增益是：

$$\mathrm{Gain}(A) = \mathrm{Entropy}(A)(s_1, s_2, \cdots, s_n) - \mathrm{Entropy}(A) \tag{3}$$

ID3 选择使Gain(A)最大的属性A*作为根节点，对A*的不同取值对应的E的v个子集 E_i 递归调用上述过程生成A*的子节点，从而生成一棵树。

使用 R 软件，利用 rpart 函数包和 rpart.plot 函数包对危险因素进行挖掘。将训练集中有关最终妊高症的诊断结果(即“是”与“否”)作为最终的分类结果(即根节点)，将患者的体检属性变量作为分类条件进行分析，将影响最终诊断的危险因素及其数据范围用决策树展现出来，并将其可视化。由于决策树分支太多，过于复杂，容易产生过拟合现象，对预测测试集数据丧失意义，因此利用 CP(Complexity Parameter)即复杂度参数进行剪枝。CP 随决策树复杂度的增加而减小，当增加一个节点引起的分类精确度的变化量小于决策树复杂度变化量的 CP 倍时，须剪去该节点。一般选择错判率最小值对应的 CP 值来修树。在 CP 值等于 0.004 8 时，获得既能够很好拟合训练集数据，又能很好预测测试集数据的决策树，而且对于危险因素来说，在最终得到的决策树中，强调了“收缩压”、“舒张压”、“空腹血糖”和“甘油三酯”这四个属性，根据决策树的路径显示：当收缩压大于 138 mmHg，同时舒张压大于 92 mmHg、甘油三酯大于 1.7 mmol/L 时，是主要危险因素；而当收缩压大于 138 mmHg，但舒张压小于 92 mmHg 时，若空腹血糖小于 5 mmol/L、舒张压大于 86 mmHg 且甘油三酯大于 2.6 mmol/L 时，也是患妊高症的危险因素。

(2) 逻辑回归模型对妊高症危险因素挖掘

逻辑回归(Logistic Regression, LR)模型中最常使用梯度下降法来获得代价函数的最小值，通过给予一定的优化条件，使方法得到更好的分类界限[14]。由于逻辑回归模型构造简单、结果方便易懂，因此在疾病防治领域有着广泛的应用，是数据挖掘方法在医学领域应用的一个典型方法。

设 P 为某事件发生的概率，取值范围为[0, 1]，1–P 为该事件不发生的概率，将 P/(1–P) 取自然对数 ln(P/(1–P))，即对 P 作 logit 转换，记为 logitP，则 logitP 的取值范围为(–∞, +∞)。以 P 为因变量，建立线性回归方程[15]：

$$\mathrm{logit}P = \alpha + \beta_1 x_1 + \cdots + \beta_m x_m \tag{4}$$

可得：

$$P = \frac{\exp(\alpha + \beta_1 x_1 + \cdots + \beta_m x_m)}{1 + \exp(\alpha + \beta_1 x_1 + \cdots + \beta_m x_m)} \tag{5}$$

该模型即为逻辑回归模型，是普通多元线性回归模型的推广，但它的误差项服从二项分布而非正态分布，模型中 α 为常数，β_i (i=1, …, m)为逻辑回归系数。

使用 R 软件，利用 glm 函数和 MASS 函数包对危险因素进行挖掘。由于逻辑回归模型没有参数，因此不需要对参数进行调整，为了得到既能很好拟合训练集，又能很好预测测试集数据的模型，需要选择合适的属性，此属性即为模型挖掘的危险因素。通过对变量进行合理选择，得到合理的逻辑回归模型及其可视化图谱。逻辑回归模型通过 8 次费舍尔得分迭代，筛

选出有意义的属性变量为“年龄”、“体重 Kg”、“收缩压 mmHg”、“舒张压 mmHg”和 “空腹血糖”, 通过这 5 个属性变量, 在 R 中建立针对妊高症诊断的最合理模型, 如果将这 5 个属性以“X_1-X_5”分别表示, Y 为患者最终是否患病的结果(Y 值只能为 0 或 1), 则笔者提出的最终逻辑回归公式可以表示为:

$$Y = -25.45 - 0.05X_1 + 0.03X_2 + 0.17X_3 + 0.01X_4 - 0.21X_5 \quad (6)$$

根据公式(6)进行计算, 以说明患者是否真正患有妊高症。通过重新建立逻辑回归模型, 使之包含“年龄”、“体重 Kg”、“收缩压 mmHg”、“舒张压 mmHg”和“空腹血糖”等 5 个属性, 不仅建立起基于训练集数据的逻辑回归模型, 也筛选出影响妊高症的危险因素。

(3) 神经网络模型对妊高症危险因素挖掘

神经网络(Neural Network)是一个包含输入层、隐藏层和输出层的数据挖掘方法, 神经网络方法的内在本质是: 结果与输入层的特征值无关, 是与隐藏层的方法密切相关的, 神经网络模型可以快速地学习任意的特征项。数据挖掘软件中通常运用反向传播方法使代价函数最小。神经网络可以运用于分类和回归问题, 具有极强的容错性和鲁棒性[16]。

神经网络中每个神经元都是一个简单的计算装置, 其特性由简单的数学函数所描述。神经元 i 接收其他神经元传递来的输入信息, 根据和函数 net_i 进行加权平均, 根据传递函数 f_i 产生输出信息, 输出信息又按照网络的拓扑结构传递到下一个神经元。笔者应用 McClelland 等于 1986 年提出的函数[17], 公式如下:

$$I_i = \sum_j w_{ij} x_j + Q_i \quad x_i' = f_i = \frac{1}{1 + e^{-net_i}} \quad (7)$$

其中, I_i 为神经元 i 的输入; x_i' 为神经元 i 的输出; w_{ij} 为神经元 i, j 之间的连接权; Q_i 为神经元 i 的偏置。

每一条连接弧都被赋予一定的数值表示连接弧的连接强度。正的权值表示影响的增加, 负的权值表示影响的减弱。在前向网络中, 神经元间前向连接, 同层神经元互不连接, 信息只能向着一个方向传播。前向网络的连接模式用权值向量 W 表示。在网络中, 权值向量决定着网络如何对环境中的任意输入做出反应。同样, 网络也是通过不断调整权值完成整个学习过程。

用神经网络挖掘算法对训练集数据进行处理时, 运用 R 中的 nnet 函数包和 mlbench 函数包。通过不断实验, 改变隐藏层数目和阈值, 不断优化神经网络模型。最后得到一个含有 10 个隐藏层, 阈值为 0.01 的神经网络模型。

(4) 妊高症危险因素挖掘结果

对于诊断妊高症来说, 危险因素起着至关重要的作用, 在本文的研究数据中, 一共包含 16 个属性, 但并不是全部属性都在某种程度上导致了妊高症的发生, 研究中通过决策树、逻辑回归和神经网络三种数据挖掘模型, 找到真正起作用的危险因素, 具体的挖掘结果如表 1 所示:

表 1　妊高症主要危险因素挖掘效果对比

对比项	决策树	逻辑回归	神经网络
是否可看出挖掘的危险因素	是	是	否
挖掘的危险因素属性	收缩压、舒张压、空腹血糖、甘油三酯	年龄、体重、收缩压、舒张压、空腹血糖	无
表现危险因素的方式	决策树路径(带有具体数值)	数学公式	无(黑盒模型)

从表 1 可以看出, 在挖掘妊高症危险因素方面, 决策树能提炼出危险因素的属性组合和数值; 逻辑回归只能分析危险因素的属性; 神经网络则无法获知属性和数值。因此决策树在妊高症危险因素挖掘中直观性最好, 且决策树运用最少的属性就可以判断出患者是否得病, 说明其代表性也最强。通过这些危险因素的挖掘, 可以对临床医生的诊断起到辅助作用, 对妊高症疾病的预防和预后起到指导作用。

4　模型评价

4.1　评价指标

大数据分析中, 利用上述三种数据挖掘模型对测试集数据进行预测, 以四格表为数据基础, 运用查准率(Precision)、召回率(Recall)、正确率和F值[18]这 4 个指标评价数据挖掘算法的性能。

各个指标具体的含义为:

$$查准率 = \frac{TP}{TP + FP} \quad (8)$$

$$召回率=\frac{TP}{TP+FN} \quad (9)$$

$$正确率=\frac{TP+TN}{TP+TN+FP+FN} \quad (10)$$

$$F值=\frac{2PR}{P+R} \quad (11)$$

在医学领域，TP(True Positive)表示真阳性的病例数，即医生诊断的结果和数据挖掘结果都是妊高症的病例数；TN(True Negative)是真阴性，即机器诊断结果不是妊高症而且医生诊断也不是的病例数；FN(False Negative)是假阴性，即机器诊断结果是妊高症，但是医生的诊断却不是的病例数；FP(False Positive)是假阳性，即机器诊断结果不是妊高症，但是医生诊断结果却是妊高症的病例数。P 是查准率，R 是召回率。

查准率越高，说明算法的敏感性越高；召回率越高，说明算法的特异性越好；正确率越高，说明算法的精确度越好；F 值越大，说明算法的总体性能越好[19]。

4.2　三种模型预测结果

利用建立好的三种数据挖掘模型和处理好的测试集数据对妊高症患病与否进行预测，利用四格表[20]数据分别计算查准率、召回率、正确率和 F 值，并对三种模型进行评价，如表 2–表 4 所示：

表 2　决策树模型预测妊高症数量结果

患者真实诊断得病与否	决策树预测患者得病与否		合计
	否	是	
否	8 608	45	8 653
是	137	171	308
合计	8 745	216	8 961

表 3　逻辑回归模型预测妊高症数量结果

患者真实诊断得病与否	逻辑回归预测患者得病与否		合计
	否	是	
否	8 611	42	8 653
是	168	140	308
合计	8 779	182	8 961

表 4　神经网络模型预测妊高症数量结果

患者真实诊断得病与否	神经网络预测患者得病与否		合计
	否	是	
否	8 631	38	8 669
是	162	130	292
合计	8 793	168	8 961

4.3　不同数据挖掘算法性能对比分析

通过对决策树、逻辑回归和神经网络三种算法在 R 中运行、建模和预测数据时所表现的不同特点，对其在 TP、FP、FN、TN、查准率、召回率、正确率、F 值方面进行对比研究，验证其应用于妊高症时的性能，为算法的选择提供依据，如表 5 所示：

表 5　三种数据挖掘算法性能指标对比

算法	TP	FP	FN	TN	查准率	召回率	正确率	F值
决策树	171	137	45	8 608	55.52%	79.71%	97.97%	0.65
逻辑回归	140	168	42	8 611	45.45%	76.92%	97.66%	0.57
神经网络	130	162	38	8 631	44.52%	77.38%	97.77%	0.57

通过表 5 可以看出，在查准率一项中，三种算法的性能比较为：决策树>逻辑回归>神经网络，这也是其敏感度排名；在召回率一项中，性能比较为：决策树>神经网络>逻辑回归，这也是其特异性排名；在正确率一项中，性能比较为：决策树>神经网络>逻辑回归，这也是其精确度排名；最后，由于查准率和召回率是一组此消彼长的评价指标，单个运用不能总体评价算法的性能，因此用 F 值对算法的综合性能进行评价，结果为：决策树>逻辑回归≈神经网络。综合以上指标可以看出，决策树的性能最好，神经网络的性能略好于逻辑回归，但相差不大。整体来看，三种监督学习算法的性能都非常强。

4.4　结果分析

从上述挖掘模型的建立和模型评价方面进行分析，认为：

(1) 在疾病危险因素研究方面，决策树能提炼出危险因素的属性组合和数值，而逻辑回归只能分析危险因素的属性列，根据公式(6)对筛选的危险因素属性进行计算得出得病与否的结论，而神经网络由于其黑盒性特征无法提供预测危险因素的可能性，因此决策树在妊高症危险因素挖掘中直观性最好。决策树运用最少的属性就可以判断出患者是否得病，说明其代表性最好。

(2) 预测妊高症发病方面，综合各指标可以得出，对于诊断、预防和预后妊高症来说，决策树的性能最好，神经网络次之，而逻辑回归最差，可能是由于逻辑回归的二分类性能和神经网络的“黑盒性”特征所致。

(3) 决策树算法运用最少的属性即可得到最优模型，因此是适合于妊娠高血压综合征危险因素挖掘及最终疾病诊断的最优算法。

5 结　语

数据挖掘算法从大数据中挖掘出有用的知识以辅助决策，已成为国际上知识发现领域最前沿的研究方向之一，将数据挖掘算法与自然语言处理、概念映射、本体论等理论和技术结合，通过数据采集、数据清洗、模型建立和模型评价 4 方面所建立起的异构数据源知识发现框架能快速实现情报的收集和分析。数据挖掘作为一个可从繁杂的信息中进行知识发现的工具，将不再局限于单纯技术层面的研究，而是越来越多与其他应用学科进行交叉融合，因此情报人员应该嵌入到学科实现嵌入式学科服务，同时从研究中还发现，不同的数据挖掘算法对于不同的知识发现有不同的效果，应该具有针对性进行选择，从而更好地对相关领域人员进行决策支持服务。

参考文献：

[1] 曾建勋，魏来. 大数据时代的情报学变革[J]. 情报学报，2015, 34(1): 37-44. (Zeng Jianxun, Wei Lai. The Changes of Information Science in Big Data Era [J]. Journal of the China Society for Scientific and Technical Information, 2015, 34(1): 37-44.)

[2] Ackoff R L. From Data to Wisdom [J]. Journal of Applies Systems Analysis, 1980(16): 3-9.

[3] Bellinger G, Castro D, Mills A. Data, Information, Knowledge, and Wisdom [EB/OL]. [2015-11-24]. http://www.systems-thinking.org/dikw/dikw.htm.

[4] Zeleny M. Human Systems Management: Integrating Knowledge, Management and Systems [M]. Singapore: World Scientific, 2005: 15-16.

[5] CIO 时代网. DIKW: 数据、信息、知识、智慧的金字塔层次体系[EB/OL]. [2014-11-24]. http: //www.ciotimes.com. (CIO Network Era. DIKW: Pyramid Hierarchy of Data, Information, Knowledge, Wisdom [EB/OL]. [2014-11-24]. http://www.ciotimes.com.)

[6] 王曰芬. 文献计量法与内容分析法综合研究的方法论来源与依据[J]. 情报理论与实践，2009, 32(2): 21-26. (Wang Yuefen. The Source and Basis of the Methodology of Synthetic Research with Bibliometric Method and Content Analysis Method [J]. Information Studies: Theory & Application, 2009, 32(2): 21-26.)

[7] 王丽伟，李梅，牟冬梅，等. 一种面向知识服务的领域知识发现流程及实例研究[J]. 情报学报，2015, 34(1): 45-52. (Wang Liwei, Li Mei, Mu Dongmei, et al. A Knowledge Service-oriented Domain Knowledge Discovery Process [J]. Journal of the China Society for Scientific and Technical Information, 2015, 34(1): 45-52.)

[8] 徐戈，王厚峰. 自然语言处理中主题模型的发展[J]. 计算机学报，2011, 34(8): 1423-1436. (Xu Ge, Wang Houfeng. The Development of Topic Models in Natural Language Processing [J]. Chinese Journal of Computers, 2011, 34(8): 1423-1436.)

[9] 何清，李宁，罗文娟，等. 大数据下的机器学习算法综述[J]. 模式识别与人工智能，2014, 27(4): 327-336. (He Qing, Li Ning, Luo Wenjuan, et al. A Survey of Machine Learning Algorithms for Big Data [J]. PR&AI, 2014, 27(4): 327-336.)

[10] 唐慧丰，谭松波，程学旗. 基于监督学习的中文情感分类技术比较研究[J]. 中文信息学报，2007, 21(6): 88-94, 108. (Tang Huifeng, Tan Songbo, Cheng Xueqi. Research on Sentiment Classification of Chinese Reviews Based on Supervised Machine Learning Techniques [J]. Journal of Chinese Information Processing, 2007, 21(6): 88-94, 108.)

[11] 侯亚君. R 语言在数据挖掘中的运用[J]. 晋城职业技术学院学报，2014, 7(2): 63-65. (Hou Yajun. On the Application of R Language in Data Mining [J]. Journal of Jincheng Institute of Technology, 2014, 7(2): 63-65.)

[12] 杨静，张楠男，李建，等. 决策树算法的研究与应用[J]. 计算机技术与发展，2010, 20(2): 114-116, 120. (Yang Jing, Zhang Nannan, Li Jian, et al. Research and Application of Decision Tree Algorithm [J]. Computer Technology and Development, 2010, 20(2): 114-116, 120.)

[13] 洪家荣，丁明峰，李星原，等. 一种新的决策树归纳学习算法[J]. 计算机学报，1995, 18(6): 470-474. (Hong Jiarong, Ding Mingfeng, Li Xingyuan, et al. A New Algorithm of Decision Tree Induction [J]. Chinese Journals of Computers, 1995, 18(6): 470-474.)

[14] 邢秋菊，赵纯勇，高克昌. 基于 GIS 的滑坡危险性逻辑回归评价研究[J]. 地理与地理信息科学，2004, 20(3): 49-51. (Xing Qiuju, Zhao Chunyong, Gao Kechang. Logical Regression Analysis on the Hazard of Landslide Based on GIS [J]. Geography and Geo-Information Science, 2004, 20(3): 49-51.)

[15] 邬伦，刘瑜，张晶，等. 地理信息系统——原理、方法和应用[M]. 北京：科学出版社，2001. (Wu Lun, Liu Yu, Zhang

Jing, et al. Geographical Information System——Theory, Method, Application [M]. Beijing: Science Press, 2001.)

[16] 王春峰, 万海晖, 张维. 基于神经网络技术的商业银行信用风险评估[J]. 系统工程理论与实践, 1999(9): 24-32. (Wang Chunfeng, Wan Haihui, Zhang Wei. Credit Risk Assessment in Commercial Banks Using Neural Networks [J]. System Engineering Theory and Practice, 1999(9): 24-32.)

[17] McClelland J L, Rumelhart D E, Hinton G E. Parallel Distributed Processing: Explorations in the Microstructure of Cognition [M]. Cambridge, MA: MIT Press, 1986.

[18] Zhang Y, Cui H, Burkell J, et al. A Machine Learning Approach for Rating the Quality of Depression Treatment Web Pages [C]. In: Proceedings of iConference 2014.

[19] Manning C D, Schutze H, Raghavan P. 信息检索导论 [M]. 王斌译. 北京: 人民邮电出版社, 2010: 105-107, 196-200. (Manning C D, Schutze H, Raghavan P. Introduction to Information Retrieval [M]. Translated by Wang Bin. Beijing: Posts & Telecom Press, 2010: 105-107, 196-200.)

[20] 赵莹. 配对四格表资料的条件 Logistic 回归模型的 Bayes 分析[J]. 数理医药学杂志, 2010, 23(5): 505-506. (Zhao Ying. Bayes Analysis of Conditional Logistic Model for Paired Fourfold Table Data [J]. Journal of Mathematical Medicine, 2010, 23(5): 505-506.)

作者贡献声明:

牟冬梅: 提出研究思路, 设计研究方案和技术路线, 论文撰写与定稿;

任珂: 研究过程的实施, 数据清洗及数据分析, 论文撰写。

利益冲突声明:

牟冬梅, 任珂在本文研究中使用了长春市妇产医院的电子病历数据。

支撑数据:

支撑数据[1]见期刊网络版 http://www.infotech.ac.cn; 支撑数据[2-3]由作者自存储, E-mail: moudm@jlu.edu.cn。

[1] 牟冬梅, 任珂. prog_code.rdf. 疾病预测模型实验环境、程序代码与结果.

[2] 牟冬梅, 任珂. trainingData.csv. 妊娠高血压预测模型训练数据集.

[3] 牟冬梅, 任珂. testingData.csv. 妊娠高血压预测模型测试数据集.

收稿日期: 2016-02-19
收修改稿日期: 2016-03-26

Discovering Knowledge from Electronic Medical Records with Three Data Mining Algorithms

Mu Dongmei[1] Ren Ke[2]
[1](School of Public Health, Jilin University, Changchun 130021, China)
[2](School of Information Management, Wuhan University, Wuhan 430072, China)

Abstract: [**Objective**] This empirical study tries to identify risk factors for diseases from the heterogeneous Electronic Medical Records (EMR). [**Methods**] First, we collected EMR with various data structures. Second, we built models to predict risk factors for diseases with the help of three algorithms (i.e., decision-making tree, logistic regression and neutral network). Finally, we compared and evaluated these models statistically. [**Results**] The Decision Tree Model achieved higher recall and precision rates than the Logistic Regression and Neural Network ones. However, there was no significant difference among them. [**Limitations**] We did not optimize the EMR's properties. [**Conclusions**] The Decision Tree Model does a better job than the Logistic Regression and Neural Network models in discovering the risk factors to predict diseases. The framework of knowledge discovery based on data mining algorithms, provides some directions for future research.

Keywords: Knowledge discovery Electronic medical record Data mining algorithms Prediction model